Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 736301, 8 pages
doi:10.1155/2010/736301

Research Article

An MCMC Algorithm for Target Estimation in Real-Time DNA Microarrays

Haris Vikalo and Mahsuni Gokdemir

Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712-0240, USA

Correspondence should be addressed to Haris Vikalo, hvikalo@ece.utexas.edu

Received 1 February 2010; Accepted 15 July 2010

Academic Editor: Harri Lähdesmäki

DNA microarrays detect the presence and quantify the amounts of nucleic acid molecules of interest. They rely on a chemical attraction between the target molecules and their Watson-Crick complements, which serve as biological sensing elements (probes). The attraction between these biomolecules leads to binding, in which probes capture target analytes. Recently developed real-time DNA microarrays are capable of observing kinetics of the binding process. They collect noisy measurements of the amount of captured molecules at discrete points in time. Molecular binding is a random process which, in this paper, is modeled by a stochastic differential equation. The target analyte quantification is posed as a parameter estimation problem, and solved using a Markov Chain Monte Carlo technique. In simulation studies where we test the robustness with respect to the measurement noise, the proposed technique significantly outperforms previously proposed methods. Moreover, the proposed approach is tested and verified on experimental data.

1. Introduction

Molecular biosensors [1] are devices that contain a biological sensing element closely coupled with a transducer. They measure interaction of biomolecules of interest (*target analytes*) with the biological sensing element, and generate signal proportional to the amount of the analyte molecules. Detection in affinity biosensors [2] relies on chemical attraction between target analytes and their molecular complements, which serve as biological sensing elements (*probes*). The attraction between these biomolecules (their *affinity* for each other) leads to binding, in which probes capture target analytes. For instance, nucleic acid probes (DNA, RNA, or synthetic oligonucleotides) capture their Watson-Crick complements, antibody probes capture antigens, cell receptor probes capture ligands, and so forth. A transducer then converts the number of complex molecular structures that are formed due to the binding into a signal. Affinity biosensors can be multiplexed, which led to the development of microarrays—arrays of affinity biosensors capable of testing a large number of analytes simultaneously. DNA microarrays [3], in particular, are capable of screening tens or even hundreds of thousands of different gene sequences at the same time, revealing critical information about the functionality of cells, effects of drugs on organisms, and so forth. Microarrays are time- and cost-efficient, and may enable exciting new applications in drug discovery, medicine, defense systems, and environmental monitoring.

Despite their enormous potential, however, microarrays have not fully met the expectations of the research community and industry. Although in principle reliable [4], their performance still leaves something to be desired [5, 6]. Today, the sensitivity, dynamic range, and resolution of DNA microarrays are limited by interference, noise, probe saturation, and other sources of errors in the analyte detection procedure. Several of these limitations stem from the fact that the molecular binding is a stochastic process, which many of the conventional affinity biosensors attempt to characterize based on a single measurement of its equilibrium, that is, by taking one sample from the steady-state distribution of the binding process. On the other hand, *real-time* DNA microarrays are capable of taking multiple temporal samples

of a binding process [7–9]. However, analyte estimation therein is typically performed using only the data collected in the equilibrium, and rarely relies on the kinetics [10].

In [11], analyte targets in real-time DNA microarrays are estimated using the temporally sampled kinetics of the binding process. However, the kinetics process there is described using a deterministic model. In this paper, we propose a comprehensive stochastic model of the binding process and state a Markov Chain Monte Carlo (MCMC) algorithm for the estimation of the target analytes. The performance of the proposed algorithm is tested on both synthetic and experimental data.

The paper is organized as follows. In Section 2, we describe the stochastic differential equation modeling the probe-target binding process. In Section 3, parameter estimation in discretely sampled diffusion processes is described, assuming noiseless data acquisition. An MCMC algorithm for the parameter estimation in the realistic noisy scenario is discussed in Section 4. Section 5 shows simulation results, while the experimental verification is provided in Section 6. Section 7 concludes the paper and outlines future work.

2. Stochastic Model

Let n_t denote the total number of analyte molecules, and let $n_c(t)$ denote the number of those that are bound to their corresponding probes at time t. For simplicity, let us assume that the number of probe molecules, n_p, is greater than n_t. Then the probability that a free analyte molecule becomes captured during the $(t, t+\Delta t)$ time interval is

$$p_b = k_1\left(1 - \frac{n_c(t)}{n_p}\right)\Delta t, \tag{1}$$

where k_1 denotes the association rate of the capturing process assuming an unlimited amount of probe molecules, and $(1 - n_c(t)/n_p)$ is the fraction of the probe molecules that are available. Assuming that the binding events are mutually independent and that n_t is large, the number of analyte molecules captured during the $(t, t+\Delta t)$ time interval follows Binomial distribution with mean $(n_t - n_c(t))p_b$ and variance $(n_t - n_c(t))p_b(1 - p_b)$. For large n_t (which in the biosensor context is certainly the case), this Binomial distribution can be approximated by a Gaussian

$$\mathcal{N}\left((n_t - n_c(t))p_b, (n_t - n_c(t))p_b(1 - p_b)\right). \tag{2}$$

Following a similar argument, it can be shown that the number of analyte molecules which are released during the $(t, t+\Delta t)$ time interval is distributed with

$$\mathcal{N}\left(n_c(t)p_r, n_c(t)p_r(1 - p_r)\right), \tag{3}$$

where $p_r = k_{-1}\Delta t$ is the probability of release of a captured analyte molecule, and where k_{-1} denotes the disassociation rate (for more details see, e.g., [12]).

Now, $n_c(t)$ is a continuous-time Markov process. Its states are discrete, but under some mild conditions [13] (the transition probabilities do not change abruptly and $n_c(t)$ is sufficiently large, both of which are readily satisfied in the biosensor context), we can describe the dynamics of $n_c(t)$ by the following stochastic differential equation (SDE)

$$n_c(t + dt) - n_c(t) = \mu(n_c, \boldsymbol{\theta}, t)dt + \sigma(n_c, \boldsymbol{\theta}, t)dW, \tag{4}$$

where $\boldsymbol{\theta} = [n_t \;\; n_p \;\; k_1 \;\; k_{-1}]$, the drift $\mu(n_c, \boldsymbol{\theta}, t)$ and diffusion $\sigma(n_c, \boldsymbol{\theta}, t)$ coefficients are given by

$$\begin{aligned}\mu(n_c, \boldsymbol{\theta}, t) &= k_1\frac{n_p - n_c}{n_p}(n_t - n_c) - k_{-1}n_c,\\ \sigma(n_c, \boldsymbol{\theta}, t) &= \left[k_1\frac{n_p - n_c}{n_p}(n_t - n_c) + k_{-1}n_c\right]^{1/2},\end{aligned} \tag{5}$$

and where W denotes the Wiener process (detailed derivation is in [12]).

Real-time DNA microarrays collect noisy observations of the temporally sampled diffusion process (4). Ultimately, we would like to use the collected observations to estimate parameters of the model $\boldsymbol{\theta}$ (including n_t, the number of target molecules). A survey of techniques for parameter estimation of discretely observed diffusion processes is given in [14]. These techniques include (i) estimating functions [15]; (ii) indirect inference and efficient method of moments [16]; (iii) Bayesian analysis and Markov Chain Monte Carlo (MCMC) methods [17–20]; (iv) analytical and numerical approximation of the likelihood function [21–23]. For Bayesian analysis and the MCMC methods, the SDE is first discretized in-sync with the measurements, using time increments equal to the sampling period of the measurements. Additional time points are introduced between the samples [24], and the corresponding values of $n_c(t)$ are treated as missing data points. The MCMC techniques [25] are then used to generate the missing data points. We should point out that MCMC techniques may be employed to estimate parameters in fairly general SDE models where the drift and diffusion coefficients are allowed to be nonlinear functions of diffusion process, or where parameters may enter into these coefficients nonlinearly. This is the case for the SDE model of real-time biosensor arrays (4).

In this paper, we rely on MCMC techniques to estimate the parameters $\boldsymbol{\theta}$ of the SDE model (4) observed at discrete points in time and subject to measurement noise. In order to derive suitable proposal densities in the MCMC algorithm, we assume that the drift and diffusion coefficients satisfy the Lipschitz and linear growth conditions

$$\left|\left(\mu(x, \boldsymbol{\theta}, t) - \mu(y, \boldsymbol{\theta}, t)\right)\right| + \left|\sigma(x, \boldsymbol{\theta}, t) - \sigma(y, \boldsymbol{\theta}, t)\right| \le C\left|x - y\right|, \tag{6}$$

$$\left|\mu(x, \boldsymbol{\theta}, t)\right|^2 + \left|\sigma(x, \boldsymbol{\theta}, t)\right|^2 \le C^2\left(1 + |x|^2\right) \tag{7}$$

for some positive constant C (see, e.g., [26]). For the sake of clarity of presentation, in the next section we first consider the noise-free case. Then, in the following section, we turn our attention to the noisy case.

3. Parameter Estimation in the Noise-Free Case

Denote the set of N observations acquired over $[0, T]$ by $\mathcal{O} = \{n_c(t_1), n_c(t_2), \ldots, n_c(t_i), \ldots, n_c(t_N)\}$, where $t_i = i\Delta t$

where Δt denotes the sampling (data acquisition) period. In principle, we may try to use the observed data to form the log-likelihood,

$$L(\boldsymbol{\theta} \mid n_c(t_1), \ldots, n_c(t_N)) = \sum_{i=1}^{N} L_i(\boldsymbol{\theta}), \tag{8}$$

where $L_i(\boldsymbol{\theta}) = \log\{p(n_c(t_i), n_c(t_{i+1}); \boldsymbol{\theta})\}$, and then find $\widehat{\boldsymbol{\theta}}$ by maximizing $L(\boldsymbol{\theta}, \cdot)$. The challenge, however, is that $p(n_c(t_i), n_c(t_{i+1}); \boldsymbol{\theta})$, a closed form expression for the transitional density between two consecutive discrete observation points is unavailable for the system in (4). Therefore, the likelihood function is often approximated via various numerical techniques [27, 28]. Here we describe the data augmentation procedure.

Consider the SDE (4) over a time interval $[0, T]$, and assume that we uniformly sample $n_c(t)$ every $\Delta t = T/N$. Therefore, we assume that the value $n_c(t_{i-1})$ at the beginning of the time interval (t_{i-1}, t_i) is known. For convenience, denote $x_i = n_c(t_i)$. Finding exact analytical expression for the transition density $p(x_i|x_{i-1}, \boldsymbol{\theta})$ appears difficult to obtain. However, if Δt is very small, we could approximate it by $p(x_i|x_{i-1}, \boldsymbol{\theta}) \sim \mathcal{N}(\mu_i, \sigma_i^2)$, where $\mathcal{N}(\mu_i, \sigma_i^2)$ denotes the normal distribution with mean μ_i and variance σ_i^2, and where

$$\begin{aligned} \mu_i &= x_{i-1} + \mu(x_{i-1}, \boldsymbol{\theta}, t_{i-1})\Delta t, \\ \sigma_i^2 &= \sigma^2(x_{i-1}, \boldsymbol{\theta}, t_{i-1})\Delta t. \end{aligned} \tag{9}$$

On the other hand, the sampling time Δt used for data acquisition is typically not sufficiently small to justify the approximation above. Therefore, we further discretize the interval (t_{i-1}, t_i) dividing it into M subintervals, where each subinterval is of the length $\Delta\tau = \Delta t/M$. Following [29], we employ the Euler-Maruyama integration scheme to generate points from a sample path of $n_c(t)$ on $(t_i, t_i + (M-1)\Delta\tau)$. [Note that $t_{i+1} = t_i + M\Delta\tau$.] Denote these points by z_j, $j = 0, 1, \ldots, M-1$, where $z_0 = x_i$. We put all these latent values between (t_{i-1}, t_i) into $\mathbf{z} = (z_1, z_2, \ldots, z_{M-1})$. The Euler-Maruyama scheme generates z_j by recursively computing

$$\begin{aligned} z_j = z_{j-1} &+ \mu\Big(z_{j-1}, \boldsymbol{\theta}, t_i + (j-1)\Delta\tau\Big)\Delta\tau \\ &+ \sigma\Big(z_{j-1}, \boldsymbol{\theta}, t_i + (j-1)\Delta\tau\Big)\Delta W_j, \end{aligned} \tag{10}$$

$j = 1, 2, \ldots, M-1$, where $\Delta W_j = W(t_i + j\tau) - W(t_i + (j-1)\tau) \sim \mathcal{N}(0, \Delta\tau)$, and where $W(0) = 0$.

Now, we can form the joint distribution of latent values $\mathbf{z}$ with x_i given x_{i-1} and $\boldsymbol{\theta}$ and then integrate out the missing values to find the transition density.

$$\begin{aligned} p(x_i \mid x_{i-1}, \boldsymbol{\theta}) &= \int p(x_i, \mathbf{z} \mid x_{i-1}, \boldsymbol{\theta}) d\mathbf{z} \\ &= \int \prod_{m=1}^{M} p(z_m \mid z_{m-1}, \boldsymbol{\theta}) d\mathbf{z} \end{aligned} \tag{11}$$

where $x_i = z_M$ and $x_{i-1} = z_0$ and we use the Markov property of diffusion process. However, this multidimensional integral is not easy to evaluate. As a standard approach, we can use Monte Carlo integration together with the importance sampler to approximate this integral:

$$\begin{aligned} p(x_i \mid x_{i-1}, \boldsymbol{\theta}) &= \int \frac{p(x_i, \mathbf{z} \mid x_{i-1}, \boldsymbol{\theta})}{q(\mathbf{z})} q(\mathbf{z}) d\mathbf{z}, \\ \widehat{p}(x_i \mid x_{i-1}, \boldsymbol{\theta}) &= \frac{1}{K} \sum_{k=1}^{K} \frac{\prod_{m=1}^{M} p(z_m \mid z_{m-1}, \boldsymbol{\theta})}{\prod_{m=1}^{M-1} q(z_m \mid z_{m-1}, \boldsymbol{\theta})}, \end{aligned} \tag{12}$$

In this equations, we are using the fact that $n_c(t)$ is a Markov process to write the joint distribution as a product of marginal distributions. We are generating K sample paths of the $n_c(t)$ on the time interval (t_{i-1}, t_i) to approximate the transition density. Now, we must construct efficient importance samplers to draw the missing samples z_j.

The importance sampler that we consider draws z_j from the Euler approximation of the SDE ([24]). Then, since the $p(z_m|z_{m-1}, \boldsymbol{\theta})$ is also approximated using this discrete model; the first $M-1$ terms of the target density $p(z_m|z_{m-1}, \boldsymbol{\theta})$ and the density of the importance sampler $q(z_m|z_{m-1}, \boldsymbol{\theta})$ are identical and cancel each other and the only remaining term is the $p(x_i|z_{M-1}, \boldsymbol{\theta})$. After this cancelation, we have the following approximate transition density:

$$\widehat{p}(x_i \mid x_{i-1}, \boldsymbol{\theta}) = \frac{1}{K} \sum_{k=1}^{K} p(x_i \mid z_{M-1}, \boldsymbol{\theta}). \tag{13}$$

Therefore, the last point obtained by the scheme (10) is z_{M-1}, which can be regarded as a sample of the process $n_c(t)$ at $t_{i-1} + (M-1)\Delta\tau$. The Euler-Maruyama integration procedure is repeated K times, generating points $z_{M-1}^1, z_{M-1}^2, \ldots, z_{M-1}^K$. The approximation converges weakly to the desired process as M increases (see, e.g., [28] and the references therein). Thus the transition density can be approximated by

$$p(x_i \mid x_{i-1}, \boldsymbol{\theta}) \approx \widehat{p}(x_i \mid x_{i-1}, \boldsymbol{\theta}) \sim \frac{1}{K} \sum_{k=1}^{K} \mathcal{N}\left(\mu_i^k, \left(\sigma_i^k\right)^2\right), \tag{14}$$

where

$$\begin{aligned} \mu_i^k &= z_{M-1}^k + \mu\Big(z_{M-1}^k, t_{i-1} + (M-1)\Delta\tau, \boldsymbol{\theta}\Big)\Delta\tau, \\ \left(\sigma_i^k\right)^2 &= \sigma^2\Big(z_{M-1}^k, t_{i-1} + (M-1)\Delta\tau, \boldsymbol{\theta}\Big)\Delta\tau, \end{aligned} \tag{15}$$

for each $k = 1, 2, \ldots, K$

To summarize, in each time interval (t_{i-1}, t_i) we perform the following steps.

(1) Starting from $z_0 = x_{i-1} = n_c(t_{i-1})$, employ the Euler-Maruyama technique (10) to generate K samples of the process $n_c(t)$ at $t = t_{i-1} + (M-1)\Delta\tau$. These samples are denoted by z_i^k, $1 \le k \le K$.

(2) Use $z_{M-1}^1, z_{M-1}^2, \ldots, z_{M-1}^K$ to estimate the transition density according to (14).

The approximate transition density converges to the true one as $K \to \infty$. We repeat the above procedure to obtain

approximate transition densities $\hat{p}(x_i|x_{i-1},\boldsymbol{\theta})$ for each $i = 1,2,\ldots,N$, and form the likelihood function

$$\hat{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \hat{p}(x_i \mid x_{i-1},\boldsymbol{\theta}). \tag{16}$$

Finally, $\hat{L}(\boldsymbol{\theta})$ is maximized over $\boldsymbol{\theta}$. For large M, K, the resulting $\hat{\boldsymbol{\theta}}$ approaches the true ML estimate of $\boldsymbol{\theta}$.

To lower the computational complexity of the approach described in this section, various modifications have been proposed. For instance, alternative importance samplers are employed to accelerate the convergence of the Monte Carlo integration, resulting in significant computational savings (see, e.g., [30] and the references therein). We shall not pursue these alternative importance samplers here. Instead, we switch our attention to the estimation problem in the noisy measurement case.

4. An MCMC Algorithm for Parameter Estimation in Noisy Case

The technique described in the previous section assumes noise-free data. In this section, we focus our attention on the more realistic noisy scenario. We do not explicitly form the likelihood function but instead rely on an MCMC technique which alternates between drawing missing data conditioned on parameters and observations, and the parameters conditioned on the missing data and the observations. Assume that the continuous diffusion (4) is sampled, and denote the obtained noisy observations by y_{iM}, that is, assume the continuous-discrete model

$$\begin{aligned} dn_c &= \mu(n_c,\boldsymbol{\theta},t)dt + \sqrt{\beta(n_c,\boldsymbol{\theta},t)}dW, \\ y_{iM} &= n_c(t_i) + v_i = n_c(iM\Delta\tau) + v_i, \end{aligned} \tag{17}$$

where v_i denotes iid Gaussian noise $\mathcal{N}(0,\epsilon^2)$, and where $\beta(n_c,\boldsymbol{\theta},t) = \sigma^2(n_c,\boldsymbol{\theta},t)$ is introduced for notational convenience. (Note that for the sake of simplicity we set the transduction coefficient in the measurement equation to 1.) Let $\mathcal{O}$ denote the set of collected noisy observations, $\mathcal{O} = \{y_0, y_M, \ldots, y_{iM}, \ldots, y_K\}$, where $K = NM$. Furthermore, we denote $z_i = n_c(i\Delta\tau)$ and collect the points z_i into $\mathcal{Z} = \{z_0, z_1, \ldots, z_M, \ldots, z_{2M}, \ldots, z_K\}$. (Note that y_{iM} is a noisy observation of z_{iM}.)

Following [19], to enable estimation of the parameters $\boldsymbol{\theta}$, we form the joint posterior density of the parameters and simulated missing data

$$p(\mathcal{Z},\boldsymbol{\theta} \mid \mathcal{O}) \propto p(\boldsymbol{\theta})p(z_0)\prod_{i=0}^{K-1} p(z_{i+1} \mid z_i,\boldsymbol{\theta})\prod_{i=0}^{N} p(y_{iM} \mid z_{iM},\boldsymbol{\theta}), \tag{18}$$

where the transition density $p(z_{i+1}|z_i,\boldsymbol{\theta})$ and the measurement density $p(y_i|z_i,\boldsymbol{\theta})$ are given by

$$\begin{aligned} p(z_{i+1} \mid z_i,\boldsymbol{\theta}) &= \mathcal{N}(z_{i+1}; z_i + \mu_i\Delta\tau, \beta_i\Delta\tau), \\ p(y_i \mid z_i,\boldsymbol{\theta}) &= \mathcal{N}(y_i; z_i, \epsilon^2) \end{aligned} \tag{19}$$

and where $\mu_i = \mu(z_i,\boldsymbol{\theta},i\Delta\tau)$, $\beta_i = \beta(z_i,\boldsymbol{\theta},i\Delta\tau)$. We rely on the Gibbs sampling technique to draw the missing data conditioned on the current state of the parameters and observations, and draw the parameters conditioned on the simulated missing data and observations. This procedure generates a Markov chain whose stationary distribution is (18). Expressed algorithmically, we perform the following steps.

(1) Initialize parameters and latent values. Use linear interpolation between the measured points in $\mathcal{O}$ to initialize $\mathcal{Z}$. Set the iteration counter to $s = 1$.

(2) In the iteration s, draw $\mathcal{Z}_s \sim p(\cdot|\boldsymbol{\theta}_{s-1},\mathcal{O})$.

(3) Draw $\boldsymbol{\theta}_s \sim p(\cdot|\mathcal{Z}_s,\mathcal{O})$ via Gaussian random walk update.

(4) Set $s = s + 1$ and go to step 2.

Finding the analytical expressions of the distributions in steps 2 and 3 appears infeasible. Hence, we employ the Metropolis-Hasting (M-H) algorithm to compute them numerically. In step 2, we generate a single component of $\mathcal{Z}$ (i.e., z_i) at a time (the so-called single site update), where there are four different cases depending on the value of the time index i. Case 1 deals with drawing the missing data z_i for which there are no corresponding noisy observations in $\mathcal{O}$ (i.e., i is not an integer multiple of M). On the other hand, Cases 2–4 deal with drawing the missing data z_i for which we do acquire noisy measurements. Among these, Cases 3 and 4 deal with the missing data at the start and at the end of the binding process, respectively (i.e., the boundary points corresponding to $i = 0$ and $i = K$). Case 2 deals with drawing the remaining missing data z_i (i.e., i is an integer multiple of M, $i \neq 0,\ K$).

Case 1. i (is not an integer multiple of M). In this case, the conditional distribution is given by

$$p(z_i \mid z_{i-1}, z_{i+1},\boldsymbol{\theta}) \propto p(z_i \mid z_{i-1},\boldsymbol{\theta})p(z_{i+1} \mid z_i,\boldsymbol{\theta}). \tag{20}$$

Direct sampling from this distribution is not feasible. Therefore, we need to employ the M-H algorithm.

Following [17], when the drift and the diffusion coefficients are constant it holds that

$$p(z_i \mid z_{i-1}, z_{i+1},\boldsymbol{\theta}) \sim \mathcal{N}\left(\frac{1}{2}(z_{i-1} + z_{i+1}), \frac{1}{2}\beta\Delta\tau\right). \tag{21}$$

However, we need to consider a more general case where drift and diffusion coefficients are functions of parameters $\boldsymbol{\theta}$ and the diffusion process z (clearly, this is the case for our model).

Now, drift and diffusion coefficients have bounded growth as stated in (7); moreover, sample paths of the diffusion process (i.e., the molecular binding process) are continuous since the sample paths of the underlying Brownian motion are continuous. This implies that the drift and diffusion coefficients are locally constant. Thus, for small time interval $\Delta\tau$ the previous result stated for constant drift and diffusion coefficients also holds for arbitrary drift and

diffusion. The rigorous proof is given in [17]. It follows that a suitable proposal density $q(z_i^* | z_{i-1}, z_{i+1}, \boldsymbol{\theta})$ is given by

$$\mathcal{N}\left(z_i^*; \frac{1}{2}(z_{i-1}+z_{i+1}), \frac{1}{2}\beta(z_{i-1},\boldsymbol{\theta})\Delta\tau\right). \tag{22}$$

It can be shown that

$$q(z_i^* \mid z_{i-1}, z_{i+1}, \boldsymbol{\theta}) \longrightarrow p(z_i \mid z_{i-1}, z_{i+1}, \boldsymbol{\theta}), \tag{23}$$

as $\Delta t \to 0$. The proposed data point

$$z_i^* \sim \mathcal{N}\left(\frac{1}{2}(z_{i-1}+z_{i+1}), \frac{1}{2}\beta(z_{i-1},\boldsymbol{\theta})\Delta\tau\right) \tag{24}$$

is accepted with probability $\min(1,\alpha)$, where

$$\alpha = \frac{p(z_i^* \mid z_{i-1},\boldsymbol{\theta})\,p(z_{i+1} \mid z_i^*,\boldsymbol{\theta})}{p(z_i \mid z_{i-1},\boldsymbol{\theta})p(z_{i+1} \mid z_i,\boldsymbol{\theta})} \times \frac{q(z_i \mid z_{i-1}, z_{i+1},\boldsymbol{\theta})}{q(z_i^* \mid z_{i-1}, z_{i+1},\boldsymbol{\theta})}. \tag{25}$$

Here, z_{i-1} is the value at the iteration s and z_{i+1} is the value obtained at iteration $s-1$ of the Gibbs Sampler.

Case 2 (i is an integer multiple of M, $i \neq 0, K$). In this case, the conditional distribution is

$$\begin{aligned} &p(z_i \mid z_{i-1}, z_{i+1}, y_i, \boldsymbol{\theta}) \\ &\quad \propto p(z_i \mid z_{i-1},\boldsymbol{\theta})p(z_{i+1} \mid z_i,\boldsymbol{\theta})p(y_i \mid z_i,\boldsymbol{\theta}). \end{aligned} \tag{26}$$

Starting from (22), we can form the joint density of z_i and y_i conditioned on z_{i-1}, z_{i+1} and as

$$\mathcal{N}\left(\frac{1}{2}\begin{pmatrix} z_{i-1}+z_{i+1} \\ z_{i-1}+z_{i+1}\end{pmatrix}, \frac{1}{2}\begin{pmatrix} \beta_{i-1}\Delta\tau & \beta_{i-1}\Delta\tau \\ \beta_{i-1}\Delta\tau & \beta_{i-1}\Delta\tau + 2\epsilon^2\end{pmatrix}\right). \tag{27}$$

From this joint Gaussian density, it is straightforward to obtain the conditional one

$$q(z_i^* \mid z_{i-1}, z_{i+1}, y_i, \boldsymbol{\theta}) \sim \mathcal{N}(z_i^*; \psi, \gamma), \tag{28}$$

where the mean and the variance are given by

$$\begin{aligned} \psi &= \frac{(z_{i-1}+z_{i+1})}{2} + \frac{\Delta\tau\beta_{i-1}(y_i - (1/2)(z_{i-1}+z_{i+1}))}{(\beta_{i-1}\Delta\tau + 2\epsilon^2)}, \\ \gamma &= \frac{\beta_{i-1}\Delta\tau}{2} - \frac{1}{2}\Delta\tau\beta_{i-1}\left(\frac{1}{2}\beta_{i-1}\Delta\tau + \epsilon^2\right)^{-1}\beta_{i-1}\frac{1}{2}\Delta\tau. \end{aligned} \tag{29}$$

The proposed value z_i^* is accepted with probability $\min(1,\alpha)$, where

$$\begin{aligned} \alpha = &\frac{p(y_i \mid z_i^*,\boldsymbol{\theta})\,p(z_i^* \mid z_{i-1},\boldsymbol{\theta})\,p(z_{i+1} \mid z_i^*,\boldsymbol{\theta})}{p(y_i \mid z_i,\boldsymbol{\theta})p(z_i \mid z_{i-1},\boldsymbol{\theta})p(z_{i+1} \mid z_i,\boldsymbol{\theta})} \\ &\times \frac{q(z_i \mid z_{i-1}, z_{i+1}, y_i,\boldsymbol{\theta})}{q(z_i^* \mid z_{i-1}, z_{i+1}, y_i,\boldsymbol{\theta})}. \end{aligned} \tag{30}$$

Case 3 ($i = 0$). The conditional distribution is given by

$$p(z_0 \mid z_1, y_0, \boldsymbol{\theta}) \propto p(z_0)p(z_1 \mid z_0,\boldsymbol{\theta})p(y_0 \mid z_0,\boldsymbol{\theta}). \tag{31}$$

Using the Euler approximation, we can write:

$$z_0 = z_1 - \mu_0\Delta\tau + \beta_0^{1/2}\Delta W. \tag{32}$$

Since sample paths of the diffusion process are continuous, and since drift and diffusion coefficients have bounded growth by assumption given in (7), μ and β are locally constant. Hence, we can approximate μ_0 by μ_1 and β_0 by β_1 which leads to

$$z_0 \approx z_1 - \mu_1\Delta\tau + \beta_1^{1/2}\Delta W. \tag{33}$$

Then,

$$p(z_0 \mid z_1,\boldsymbol{\theta}) \sim \mathcal{N}(z_0; z_1 - \mu_1\Delta\tau, \beta_1\Delta\tau). \tag{34}$$

Combining this density with the measurement error density given by

$$p(y_0 \mid z_0,\boldsymbol{\theta}) = \mathcal{N}(y_0; x_0, \epsilon^2), \tag{35}$$

we obtain the joint density of y_0 and z_0 conditioned on z_1 and $\boldsymbol{\theta}$ as

$$\begin{pmatrix} z_0 \\ y_0\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} z_1 - \mu_1\Delta\tau \\ z_1 - \mu_1\Delta\tau\end{pmatrix}, \begin{pmatrix} \beta_1\Delta\tau & \beta_1\Delta\tau \\ \beta_1\Delta\tau & \beta_1\Delta\tau + \epsilon^2\end{pmatrix}\right). \tag{36}$$

By applying the relation between the joint Gaussian distribution and its corresponding conditionals, we arrive to a suitable proposal density for the M-H algorithm given by

$$q(z_0^* \mid z_1, y_0, \boldsymbol{\theta}) \sim \mathcal{N}(z_0^*; \psi, \gamma), \tag{37}$$

where the mean and the variance are defined as

$$\begin{aligned} \psi &= z_1 - \mu_1\Delta\tau + \frac{\Delta\tau\beta_1(y_0 - (z_1 - \mu_1\Delta\tau))}{(\beta_1\Delta\tau + \epsilon^2)}, \\ \gamma &= \beta_1\Delta\tau - \Delta\tau\beta_1(\beta_1\Delta\tau + \epsilon^2)^{-1}\beta_1\Delta\tau. \end{aligned} \tag{38}$$

The proposed value z_0^* is chosen with probability $\min(1,\alpha)$, where

$$\alpha = \frac{p(z_0^*)\,p(y_0 \mid z_0^*,\boldsymbol{\theta})\,p(z_1 \mid z_0^*,\boldsymbol{\theta})}{p(z_0)p(y_0 \mid z_0,\boldsymbol{\theta})\,p(z_1 \mid z_0,\boldsymbol{\theta})} \times \frac{q(z_0 \mid z_1, y_0,\boldsymbol{\theta})}{q(z_0^* \mid z_1, y_0,\boldsymbol{\theta})}. \tag{39}$$

Case 4 ($i = K$). Now the conditional distribution is

$$p(z_K \mid z_{K-1}, y_K, \boldsymbol{\theta}) \propto p(z_K \mid z_{K-1},\boldsymbol{\theta})p(y_K \mid z_K,\boldsymbol{\theta}) \tag{40}$$

By using the Euler transition density $p(z_k|z_k - 1,\boldsymbol{\theta})$ and the measurement error density $p(y_k|z_k,\boldsymbol{\theta})$, we can form the joint density of z_k and y_k conditioned on z_{k-1} and $\boldsymbol{\theta}$ as

$$\begin{pmatrix} z_K \\ y_K\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \mu\end{pmatrix}, \begin{pmatrix} \beta_{K-1}\Delta\tau & \beta_{K-1}\Delta\tau \\ \beta_{K-1}\Delta\tau & \beta_{K-1}\Delta\tau + \epsilon^2\end{pmatrix}\right), \tag{41}$$

where $\mu = z_{K-1} - \mu_{K-1}\Delta\tau$. It follows that

$$p(z_K \mid z_{K-1}, y_K, \boldsymbol{\theta}) \sim \mathcal{N}(z_K; \psi, \gamma), \tag{42}$$

where the mean and the variance are given by

$$\psi = z_{K-1} + \mu_{K-1}\Delta\tau + \frac{\Delta\tau\beta_{K-1}(y_K - (z_{K-1} + \mu_{K-1}\Delta\tau))}{(\beta_{K-1}\Delta\tau + \epsilon^2)},$$
$$\gamma = \beta_{K-1}\Delta\tau - \Delta\tau\beta_{K-1}(\beta_{K-1}\Delta\tau + \epsilon^2)^{-1}\beta_{K-1}\Delta\tau. \tag{43}$$

In this case, we can directly sample from the above density, so there is no need for the M-H algorithm.

On another note, in step 3 of the Gibbs sampling algorithm we update $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_s^* = \boldsymbol{\theta}_s + \Omega$, where $\Omega \sim \mathcal{N}(0, \Gamma)$ and $\Gamma = \text{diag}(\gamma_i)$. (These variances determine the mixing properties of the generated Markov chain.) We again use the M-H algorithm and accept $\boldsymbol{\theta}_s^*$ with probability $\min(1, \alpha)$, where

$$\alpha = \frac{L(\boldsymbol{\theta}^* \mid \mathcal{Z}, \mathcal{O})}{L(\boldsymbol{\theta} \mid \mathcal{Z}, \mathcal{O})} = \frac{\left[\prod_{i=1}^{K-1} p(z_{i+1} \mid z_i, \boldsymbol{\theta}^*)\right]\left[\prod_{i=0}^{N} p(y_{iM} \mid z_{iM}, \boldsymbol{\theta}^*)\right]}{\left[\prod_{i=1}^{K-1} p(z_{i+1} \mid z_i, \boldsymbol{\theta})\right]\left[\prod_{i=0}^{N} p(y_{iM} \mid z_{iM}, \boldsymbol{\theta})\right]}. \tag{44}$$

When the noise variance is known, $p(y|z, \boldsymbol{\theta})$ is independent of the parameters $\boldsymbol{\theta}$ and thus we can simplify α to

$$\alpha = \frac{\left[\prod_{i=1}^{K-1} p(z_{i+1} \mid z_i, \boldsymbol{\theta}^*)\right]}{\left[\prod_{i=1}^{K-1} p(z_{i+1} \mid z_i, \boldsymbol{\theta})\right]}. \tag{45}$$

5. Simulation Results

We simulate the reaction (4), where the parameters are set to $n_p = 10^5$, $n_t = 10^3$, $k_1 = 10^{-3}$, and $k_{-1} = 10^{-3}$. The signal is sampled ($N = 300$), where the samples are perturbed by an additive Gaussian noise (zero-mean, variance ε^2). In Figure 1, we compare the square root of the relative mean-square error, $\sqrt{E\{(n_t - \hat{n}_t)^2/n_t^2\}}$, of the MCMC algorithm for stochastically modeled real-time microarrays and the least-mean-squares estimation approach for deterministically modeled (by means of ordinary differential equations) real-time microarrays (see [11] for details). (We assume that all parameters other than n_t are known.) The error is plotted as a function of the observation noise variance (the error is averaged over 100 trials). The simulation results indicate that the proposed approach significantly outperforms the least-mean-squares method over the broad range of parameters. The Gibbs Sampler is performed with $M = 5$ and $K = 1500$. The burn-in period of the algorithm is 500 iterations, while no more than 300-400 iterations are needed for the convergence (see Figure 2).

6. Experimental Verification

To verify the proposed approach in experiments, we used the real-time microarray data reported in [11]. In those experiments, cDNA targets were generated from The RNA Spikes, a commercially available set of 8 purified *Escherichia*

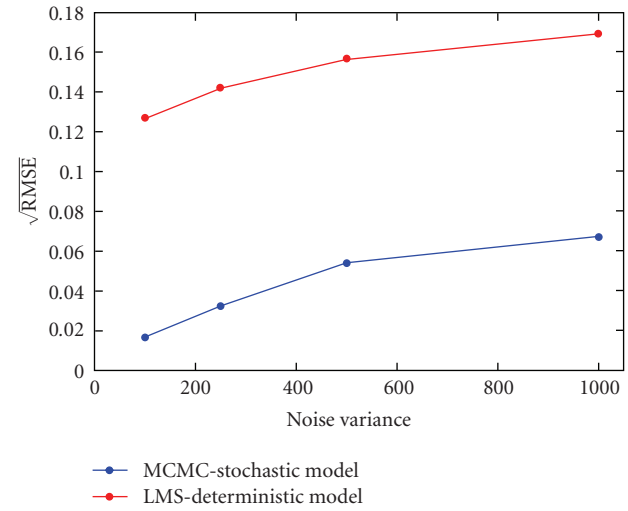

FIGURE 1: The square root of the relative mean-square error, $\sqrt{E\{(n_t - \hat{n}_t)^2/n_t^2\}}$, of the Gibbs Sampler and the least-mean-squares estimation approach, as a function of the observation noise variance of 100, 250, 500 and 1000.

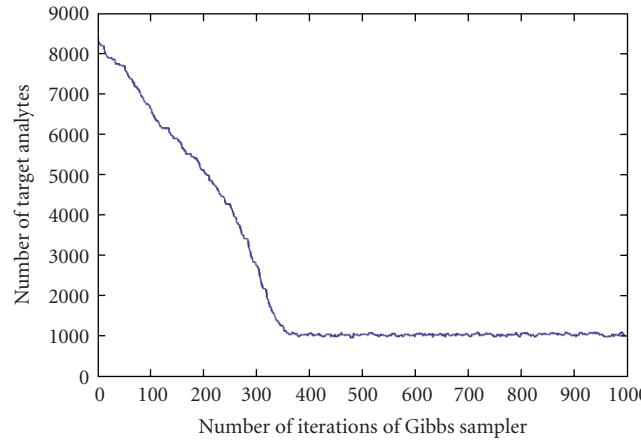

FIGURE 2: The convergence of $\hat{n}_t$ as a function of the number of iterations.

Coli RNA transcripts purchased from Ambion Inc. Lengths of the RNA sequences in the set are (750, 752, 1000, 1000, 1034, 1250, 1475, 2000), respectively. The RNA sequences were reverse transcribed to obtain the cDNA targets, which were then labeled with Cy5 dyes. Eight probes (25 mer oligonucleotides) were designed and printed on slides, where each probe was repeated in 6 different spots; hence, the printed slides had 48 spots. We focus on two experiments, one where the concentrations of the targets was 80 ng/50 μL, and the other where the concentrations of the targets was 16 ng/50 μl.

In order to mitigate the numerical problems caused by large numbers (e.g., n_p is on the order of 10^{11} in the experimental data), we scale down the variables in the SDE (in particular, scaling factor $k = 10^6$ was chosen). Then,

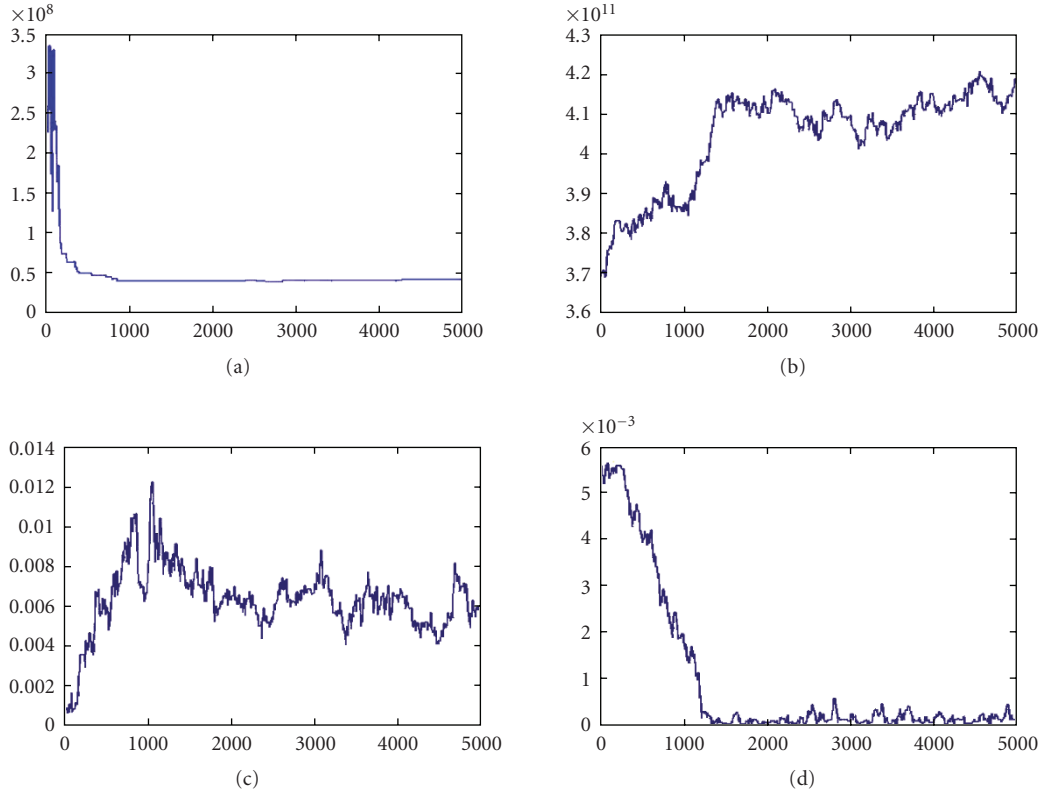

Figure 3: The convergence of parameter estimates $\hat{n}_t$ (a), $\hat{n}_p$ (b), $\hat{k}_1$ (c), and $\hat{k}_{-1}$ (d) as a function of the number of iterations of the Gibbs sampler.

exploiting the linearity of our SDE model, the scaled down continuous-discrete model is given by

$$d\overline{n}_c(t) = \overline{\mu}(n_c, \boldsymbol{\theta}, t)dt + \overline{\sigma}(n_c, \boldsymbol{\theta}, t)dW,$$
$$\overline{y}(t) = \overline{n}_c(t) + \overline{v}(t), \tag{46}$$

where $\overline{n}_c = n_c/k$, $\overline{v} = v/k$, $\overline{y} = y/k$, and

$$\overline{\mu} = k_1 \frac{\overline{n}_p - \overline{n}_c}{\overline{n}_p}(\overline{n}_t - \overline{n}_c) - k_{-1}\overline{n}_c,$$
$$\overline{\sigma} = \left[k_1 \frac{\overline{n}_p - \overline{n}_c}{\overline{n}_p}(\overline{n}_t - \overline{n}_c) + k_{-1}\overline{n}_c \right]^{1/2}, \tag{47}$$

where $\overline{n}_p = n_p/k$ and $\overline{n}_t = n_t/k$.

Moreover, since the noise variance is generally unknown, we add it to the vector of the unknown parameters, that is, $\boldsymbol{\theta} = [n_t \;\; n_p \;\; k_1 \;\; k_{-1} \;\; \epsilon]$. This requires slight modification in the step 3 of the MCMC algorithm (as described previously).

We applied the proposed Gibbs Sampler to the estimation of the parameters of the process which generated the described experimental data. We run 5000 iterations of the algorithm with $M = 5$ and $K = 1500$, and averaged its performance over 50 trials. For the first experiment, we obtained $\hat{n}_{t,1} = 3.3 \times 10^8$, while in the second experiment $\hat{n}_{t,2} = 1.1 \times 10^8$. Figure 3 and 4 show a sample convergence of the parameter estimates as a function of the number of iterations of the Gibbs sampler applied to the second data set. We note that the ratio of the estimated target amounts is $\hat{n}_{t,1}/\hat{n}_{t,2} = 3$, which is close to the true ratio of $80ng/16ng = 5$. On the other hand, unable to observe the kinetics of the hybridization process conventional microarrays would simply estimate the ratio of target molecules by the ratio of captured molecules at the end of the experimental run. For the experiments under consideration, this ratio is 2.77.

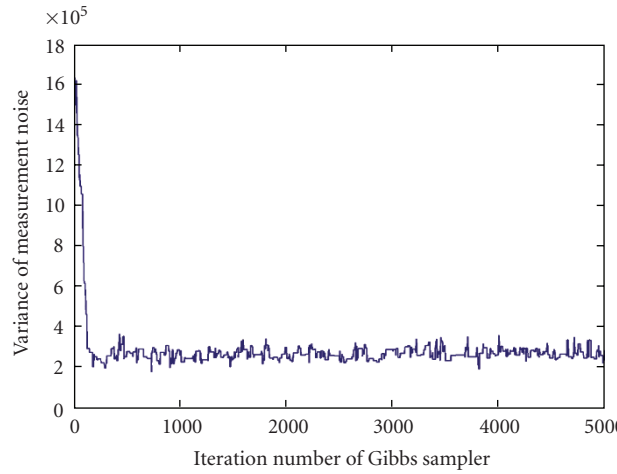

Figure 4: The convergence of the noise variance estimate as a function of the number of iterations of the Gibbs sampler.

7. Summary and Conclusion

In this paper, we considered the problem of estimating the number of target molecules in stochastically modeled biomolecular sensors. We posed it as a parameter estimation problem in systems modeled by stochastic differential equations, where the noise-perturbed data is acquired at discrete points in time. Since the problem is analytically intractable, we employed MCMC techniques to obtain a numerical solution. In particular, we relied on the use of the Gibbs Sampler to alternate between drawing missing data conditioned on parameters and observations, and drawing parameters conditioned on the simulated missing data and the observations. We used the Metropolis-Hastings technique within the Gibbs Sampler to simulate analytically untractable densities. Simulation results indicate that the proposed algorithm significantly outperforms the existing least-mean-squares approach, and that the algorithm is robust with respect to the measurement noise. Moreover, we applied the algorithm to experimental data to verify the validity of the estimation algorithm in a realistic scenario.

There are several possible extensions of the current work. For instance, the MCMC algorithm described in this paper can also be applied to multivariate diffusion processes. Such processes arise in the context of gene regulatory network as well as in real-time biosensor arrays affected by cross-hybridization. For this scenario, one may extend the algorithm so that it handles unobserved parts of a multivariate diffusion process. On another note, a variation of the MCMC algorithm performs (random) block updating (see, e.g., [19, 20]). It is worth pursuing this modification in the context of parameter estimation in real-time biosensors.

References

[1] J. Cooper and T. Cass, Eds., *Biosensors*, Oxford University Press, Oxford, UK, 2nd edition, 2004.

[2] K. R. Rogers and A. Mulchandani, *Affinity Biosensors*, Humana Press, Totowa, NJ, USA, 1998.

[3] K. R. M. Schena, *Microarray Analysis*, John Wiley & Sons, New York Ny, USA, 2003.

[4] L. Shi, L. H. Reid, W. D. Jones et al., "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, no. 9, pp. 1151–1161, 2006.

[5] E. Marshall, "Getting the noise out of gene arrays," *Science*, vol. 306, no. 5696, pp. 630–631, 2004.

[6] S. Draghici, P. Khatri, A. C. Eklund, and Z. Szallasi, "Reliability and reproducibility issues in DNA microarray measurements," *Trends in Genetics*, vol. 22, no. 2, pp. 101–109, 2006.

[7] D. I. Stimpson, J. V. Hoijer, W. Hsieh et al., "Real-time detection of DNA hybridization and melting on oligonucleotide arrays by using optical wave guides," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 14, pp. 6379–6383, 1995.

[8] J. Bishop, A. M. Chagovetz, and S. Blair, "Kinetics of multiplex hybridization: mechanisms and implications," *Biophysical Journal*, vol. 94, no. 5, pp. 1726–1734, 2008.

[9] M. R. Henry, P. W. Stevens, J. Sun, and D. M. Kelso, "Real-time measurements of DNA hybridization on microparticles with fluorescence resonance energy transfer," *Analytical Biochemistry*, vol. 276, no. 2, pp. 204–214, 1999.

[10] V. M. Mirsky, "Affinity sensors in non-equilibrium conditions: highly selective chemosensing by means of low selective chemosensors," *Sensors*, vol. 1, no. 1, pp. 13–17, 2001.

[11] H. Vikalo, B. Hassibi, and A. Hassibi, "Modeling and estimation for real-time microarrays," *IEEE Journal on Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 286–296, 2008.

[12] S. Das, H. Vikalo, and A. Hassibi, "On scaling laws of biosensors: a stochastic approach," *Journal of Applied Physics*, vol. 105, no. 10, Article ID 102021, 2009.

[13] E. Allen, *Modeling with Itô Stochastic Differential Equations*, Springer, New York, NY, USA, 2007.

[14] H. Sørensen, "Parametric inference for diffusion processes observed at discrete points in time: a survey," *International Statistical Review*, vol. 72, no. 3, pp. 337–354, 2004.

[15] B. M. Bibby, "Estimating functions for discretely sampled diffusion type models," in *Handbook of Financial Econometrics*, North-Holland, Amsterdam, The Netherlands, 2002.

[16] A. R. Gallant and J. R. Long, "Estimating stochastic differential equations efficiently by minimum chi-squared," *Biometrika*, vol. 84, no. 1, pp. 125–141, 1997.

[17] B. Eraker, "MCMC analysis of diffusion models with application to finance," *Journal of Business and Economic Statistics*, vol. 19, no. 2, pp. 177–191, 2001.

[18] G. O. Roberts and O. Stramer, "On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm," *Biometrika*, vol. 88, no. 3, pp. 603–621, 2001.

[19] A. Golightly, *Bayesian inference for nonlinear multivariate diffusion processes*, Ph.D. thesis, Newcastle University, Newcastle, UK, 2006.

[20] O. Elerian, S. Chib, and N. Shephard, "Likelihood inference for discretely observed nonlinear diffusions," *Econometrica*, vol. 69, no. 4, pp. 959–993, 2001.

[21] Y. Aït-Sahalia, "Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach," *Econometrica*, vol. 70, no. 1, pp. 223–262, 2002.

[22] A. W. Lo, "Maximum likelihood estimation of generalized Ito processes with discretely sampled data," *Econometric Theory*, vol. 4, pp. 231–247, 1988.

[23] R. Poulsen, "Approximate maximum likelihood estimation of discretely observed diffusion processes," Tech. Rep. 29, Centre for Analytical Finance, University of Aarhus, 1999.

[24] A. R. Pedersen, "A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations," *Scandinavian Journal of Statistics*, vol. 22, pp. 55–71, 1995.

[25] J. C. Spall, "Estimation via Markov chain Monte Carlo," *IEEE Control Systems Magazine*, vol. 23, no. 2, pp. 34–45, 2003.

[26] B. Oksendal, *Stochastic Differential Equations: An Introduction with Applications*, Springer, Berlin, Germany, 2003.

[27] J. P. N. Bishwal, *Parameter Estimation in Stochastic Differential Equations*, Springer, New York, NY, USA, 2007.

[28] P. Kloeden and E. Platen, *Numeric Solutions of Stochastic Differential Equations*, Springer, New York, NY, USA, 1992.

[29] M. W. Brandt and P. Santa-Clara, "Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets," *Journal of Financial Economics*, vol. 63, no. 2, pp. 161–210, 2002.

[30] G. B. Durham and A. R. Gallant, "Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes," *Journal of Business and Economic Statistics*, vol. 20, no. 3, pp. 297–316, 2002.